

Computational Human Genetics

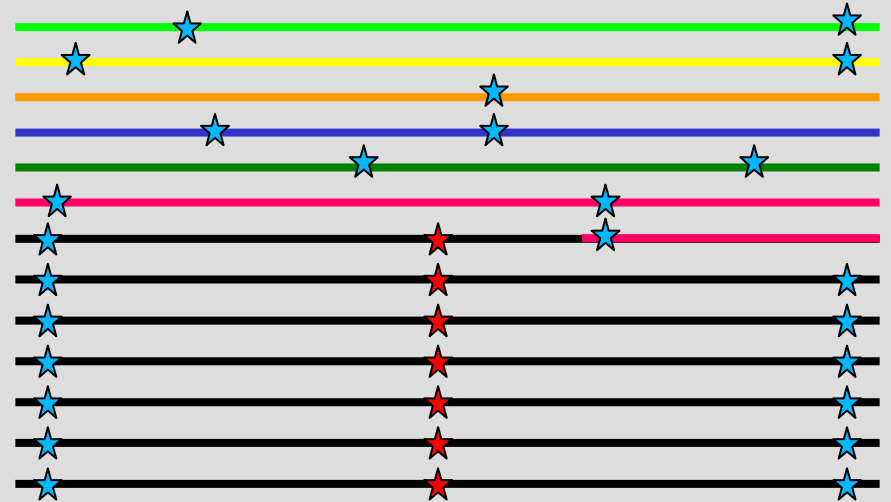
Itsik Pe'er

Department of Computer Science
Columbia University

Fall 2006

Reminder

- SNP alleles can have phenotypic outcomes:
 - Association
 - Selection



What about variants that aren't SNPs?

Meeting #9

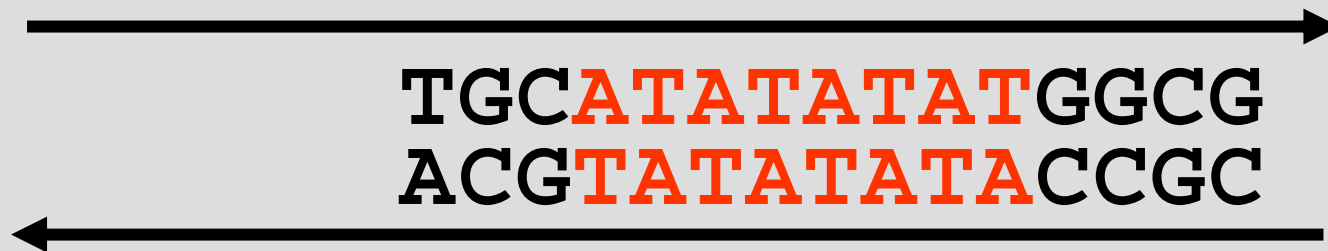
Structural Variants

Structural Variants

- Microsatellites
- Polymorphic inversions
- Polymorphic copy numbers
 - Direct methods
 - Indirect methods
- Somatic variants

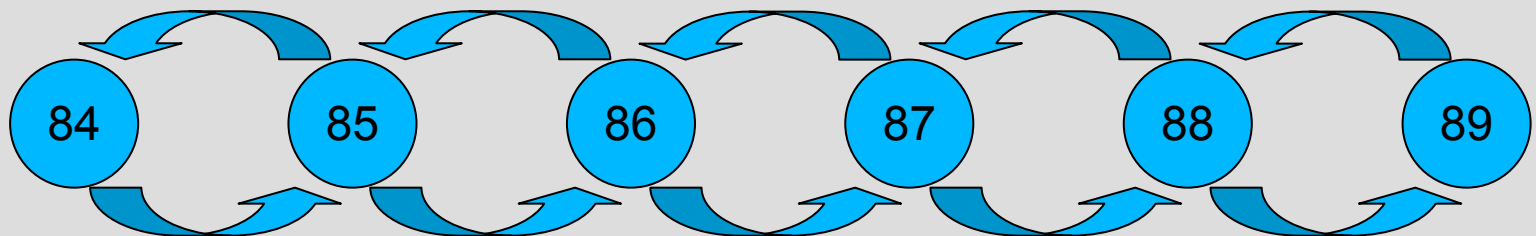
Short Tandem Repeats

- 3% of the genome ; 1/2kb



Microsatellites

- Shorter unit than minisatellites (1 to ~10)
- Most common: $(AC)_n$; $(AT)_n$
- High mutation rate (often $>10^{-3}$)
- Models of variation:
 - Many possible alleles



Microsatellite Applications

- Historically:
 - maps
- Later on:
 - Rarer alleles, linkage
- Today:
 - Combining with SNPs

Inference of Offspring Genotypes

2₁₁₀₀₁₁**2**₁₁₀₀₁₁**4**₁₁₀₀₁₁**8**₁₁₀₀**7**

5₁₀₀₁₁₀**2**₀₀₀₁₁**3**₀₁₀₀₁**6**₀₀₀₀₀**8**

6₁₀₀₁₁₀**4**₁₁₀₁₁₁**5**₁₁₀₁₁₁**3**₁₀₁₀**2**

2₁₀₁₁₁₀**3**₁₁₀₁₁₁**4**₁₁₀₀₁₁**7**₁₀₁₁**8**

5₁₀₀₁₁₀**2**₀₀₀₁₁**3**₀₁₀₀₁**6**₀₀₀₀₀**8**

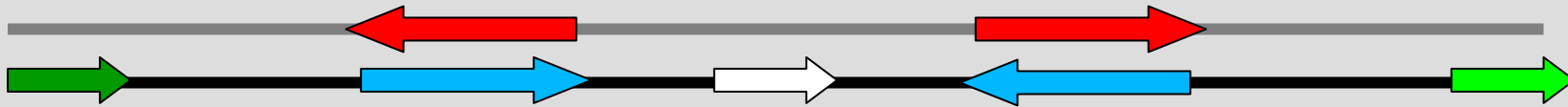
6₁₀₀₁₁₀**4**₁₁₀₁₁₁**5**₁₁₀₁₁₁**3**₁₀₁₀**2**

Structural Variants

- Microsatellites
- Polymorphic inversions
- Polymorphic copy numbers
 - Direct methods
 - Indirect methods
- Somatic variants

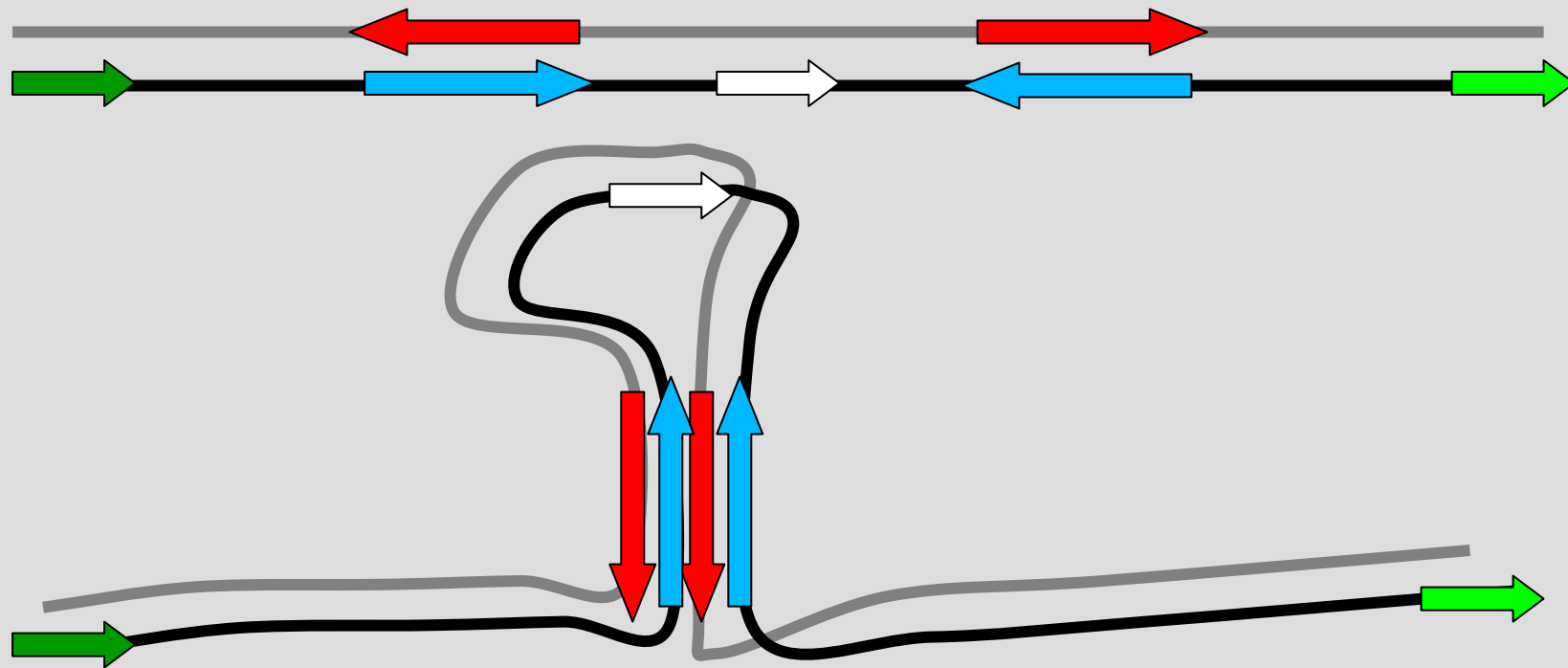
Inversions

- Result of a rare event:
non homologous recombination

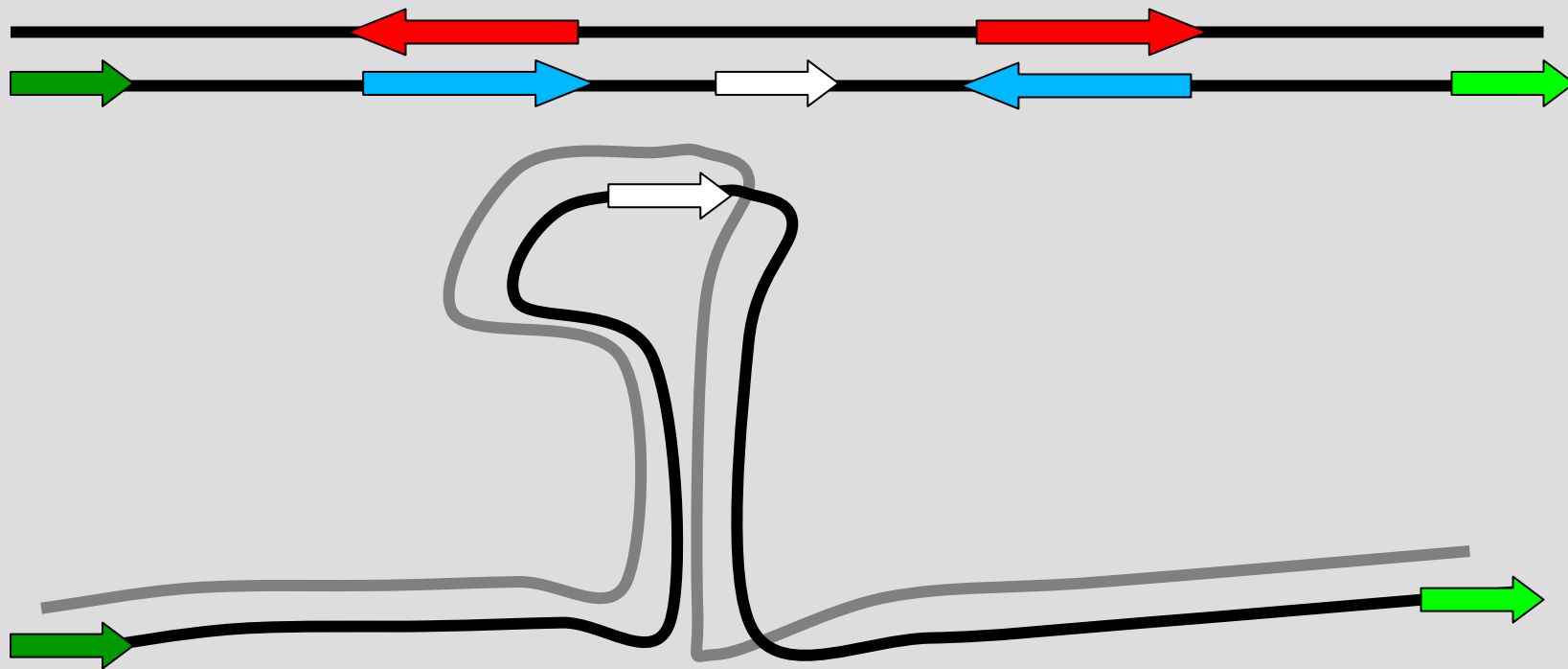


- >3 million interspersed repeats of 10^2 - 10^4 (transposons) take up 45% of the genome

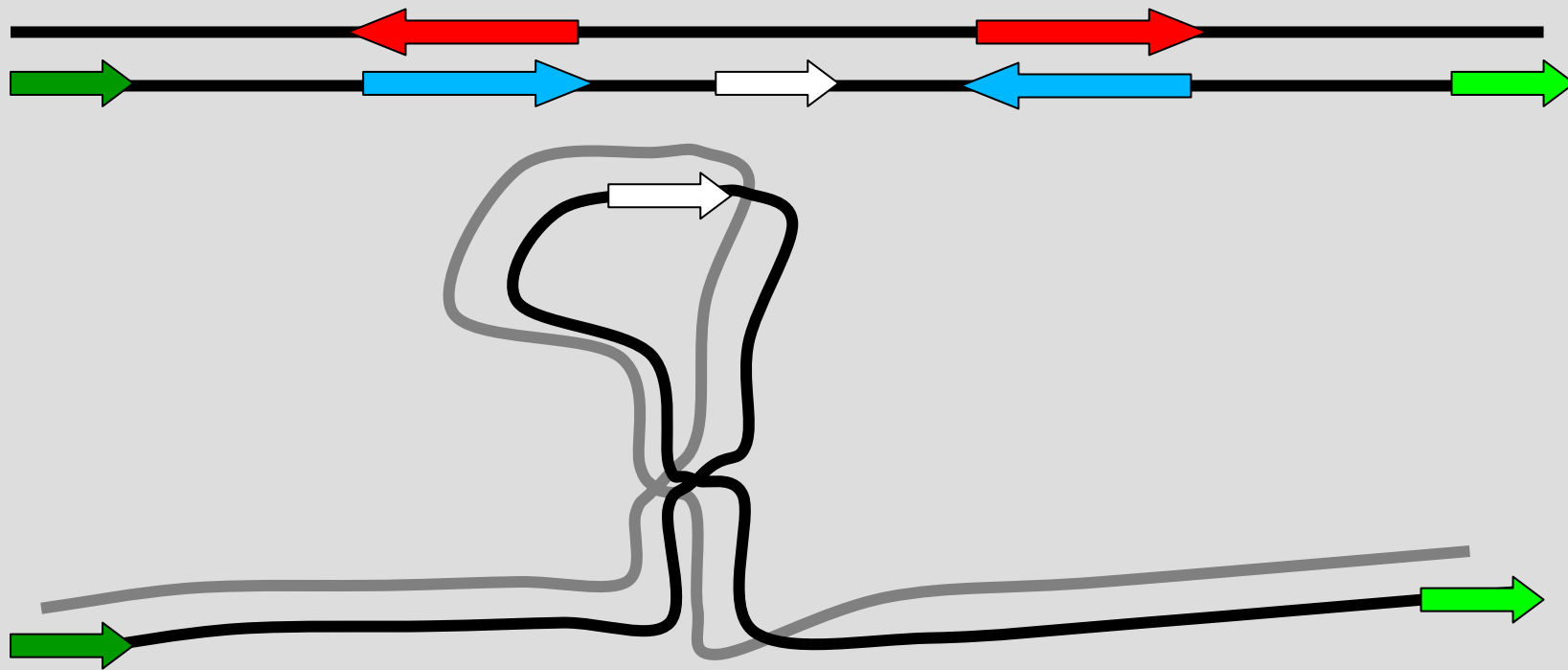
Non-homologous Recombination



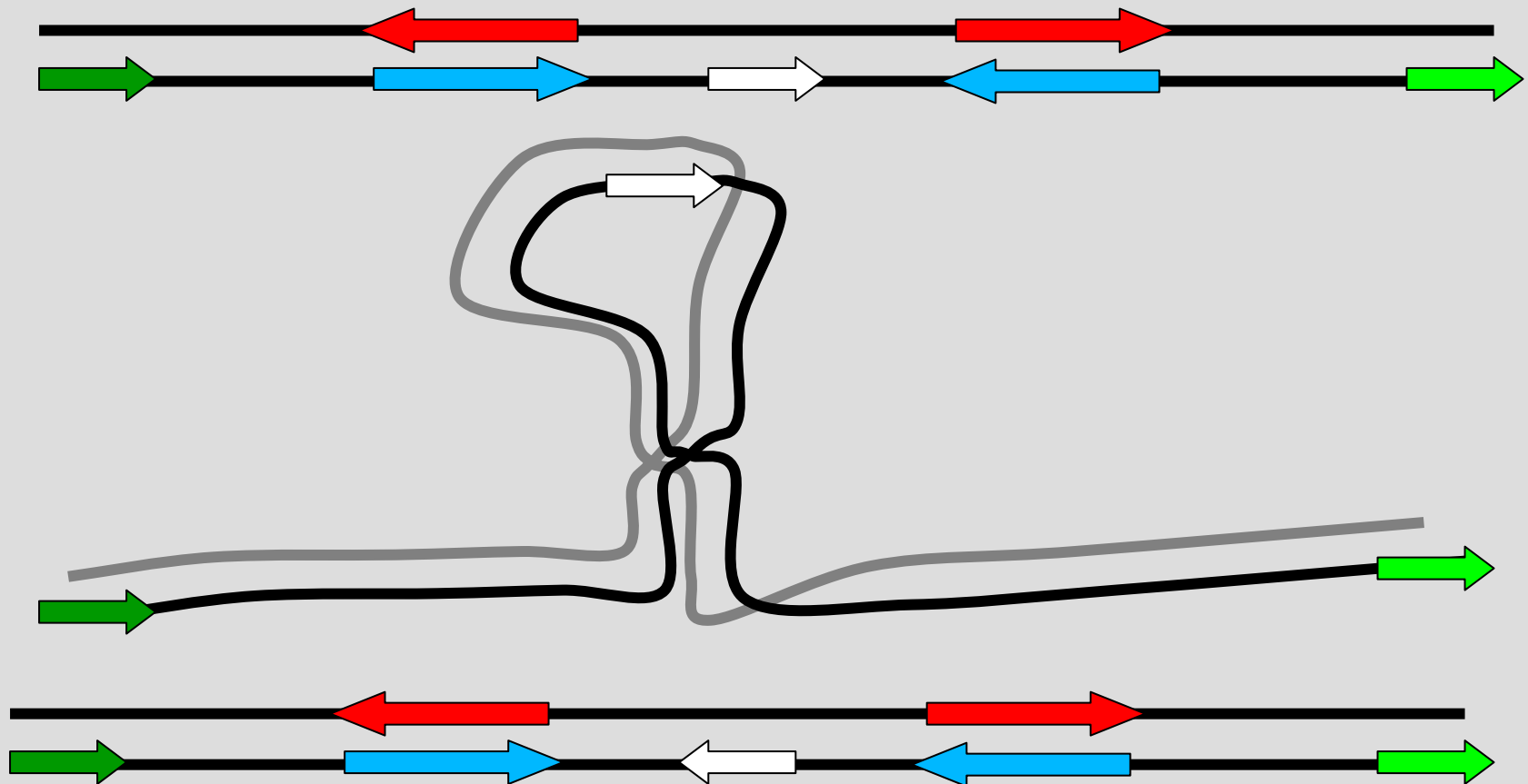
Non-homologous Recombination



Non-homologous Recombination

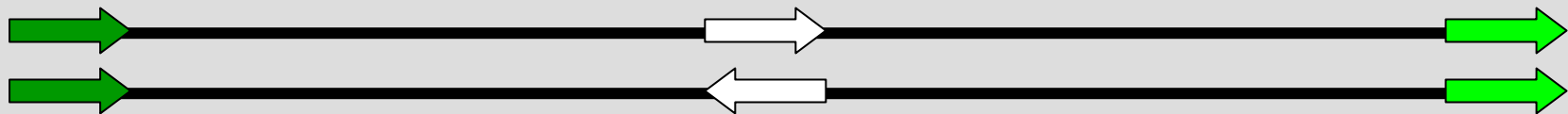


Non-homologous Recombination

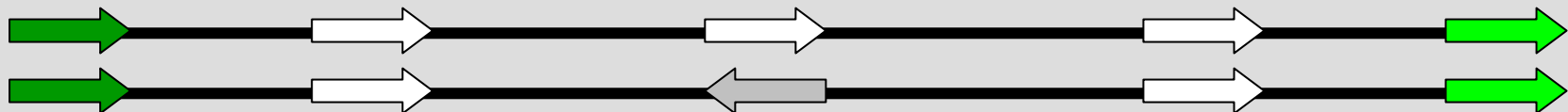


Effects of Inversions

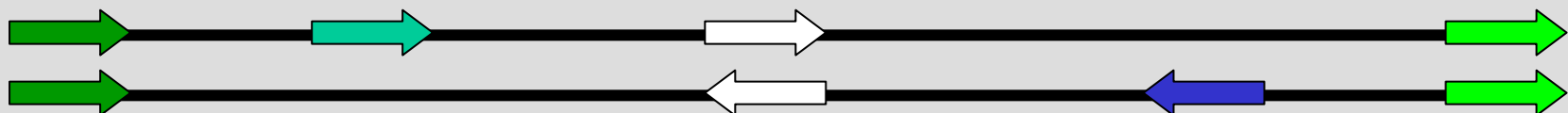
- Complete gene:
 - Typically undisrupted



- Within a gene:
 - Potential exon skipping

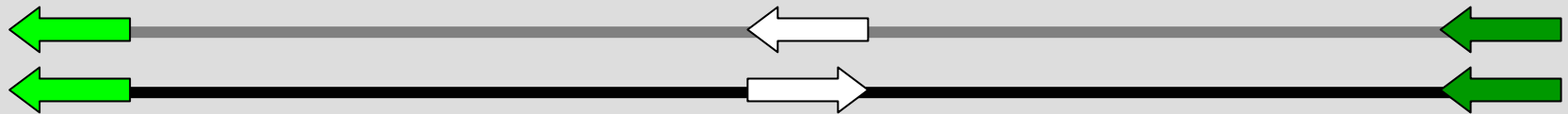


- Regulatory code:
 - Potentially replaced

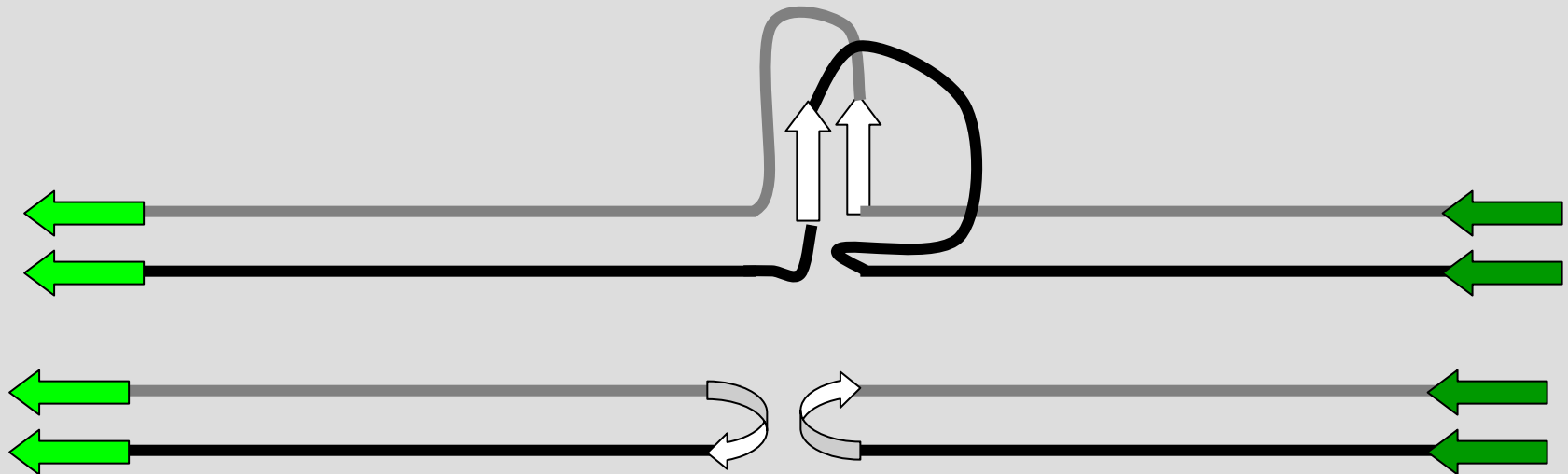


Polymorphic Inversions

- Two alleles exist; heterozygotes expected:



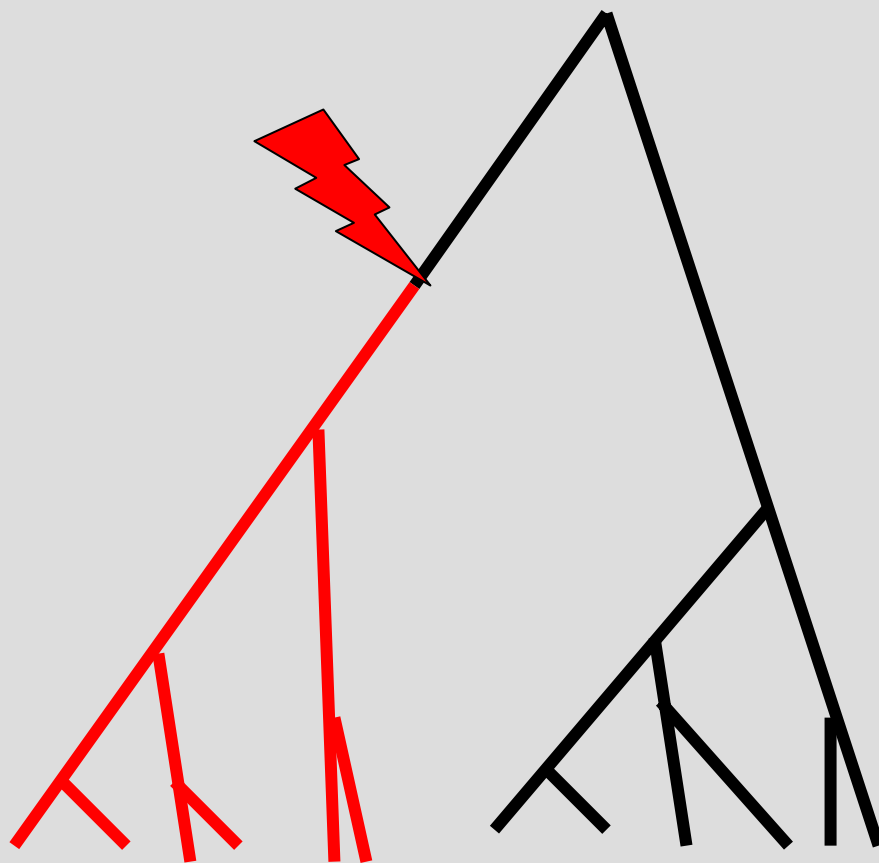
- Homologous recombination in the inversion:
Creates two same halves of a chromosome



Selection Against Nonfunctional Inversions

- Heterozygote genotypes are deleterious
- $s_h = -rd$
- In practice, selection against derived alleles
- 900kb flipped in 20% of Europeans:
 - Under + selection?

LD at Polymorphic Inversions



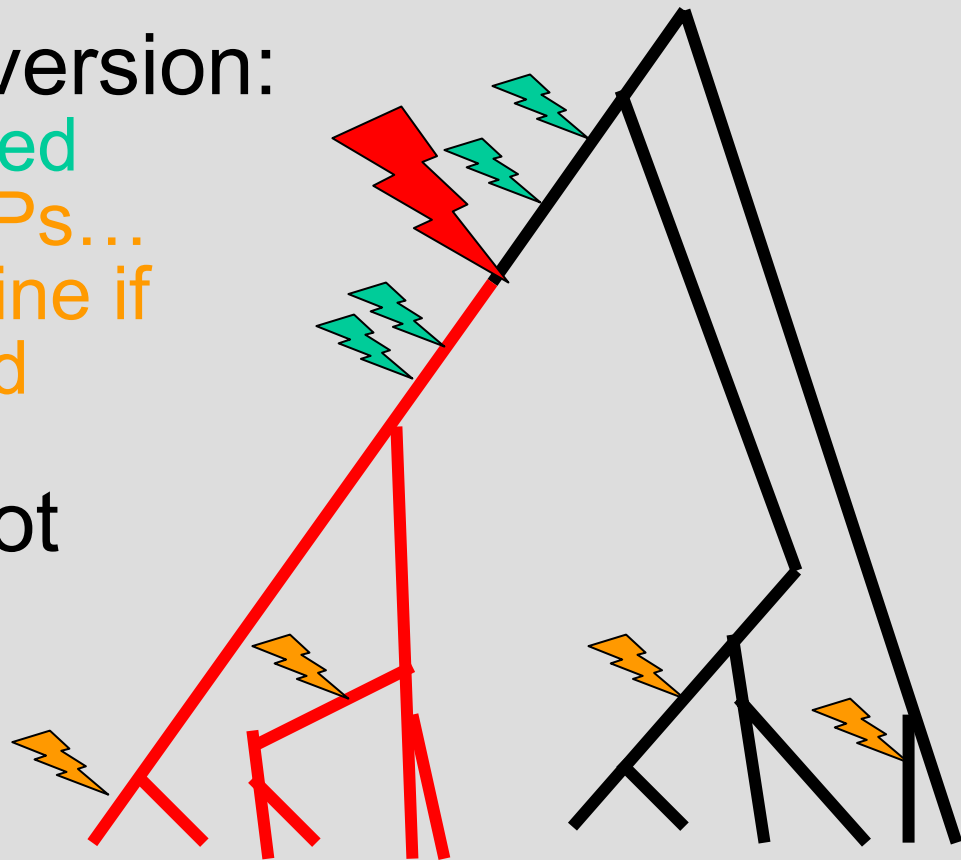
LD at Polymorphic Inversions

- Recombination may only occur between two carriers/noncarriers



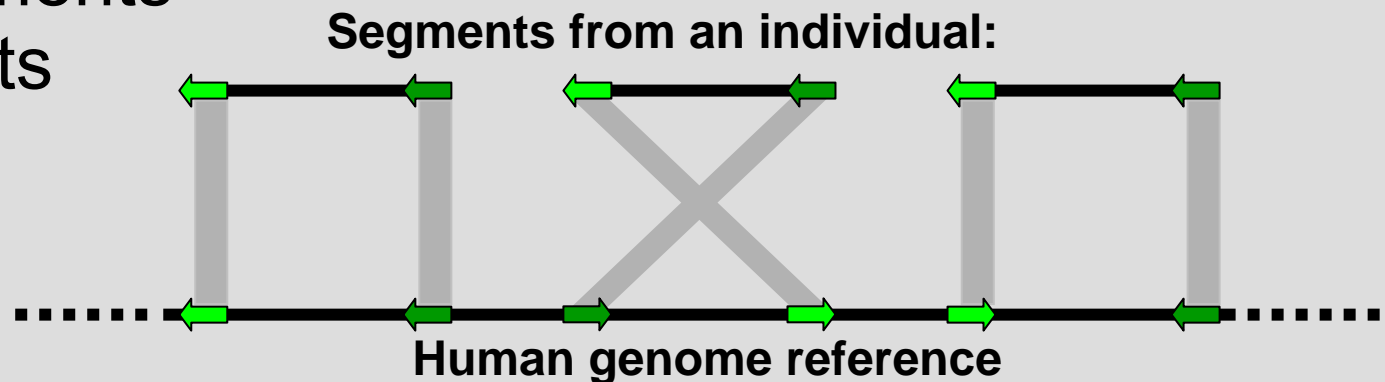
LD at Polymorphic Inversions

- Recombination may only occur between two carriers/noncarriers
- SNPs may tag inversion:
 - Perfectly correlated
 - In LD with all SNPs...
 - that may recombine if same background
- May be a blindspot for variation discovery



Experimental Survey of Inversions

- ~50 inversions found, few kb→2Mb
- Majority:
 - In large segmental duplications (5% of the genome)
- Weaknesses:
 - Short segments
 - Rare events
 - In repeats



Structural Variants

- Microsatellites
- Polymorphic inversions
- Polymorphic copy numbers
 - Direct methods
 - Indirect methods
- Somatic variants

Copy Number Variants

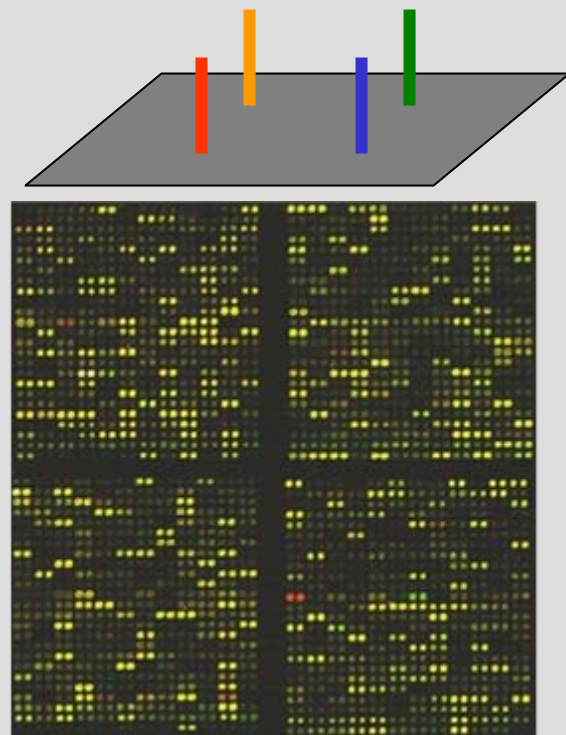
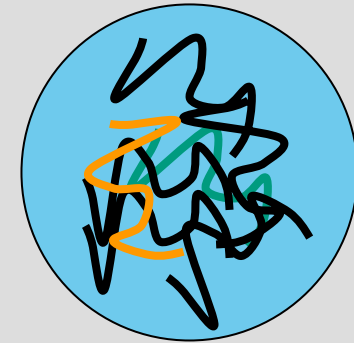


Why Should We Expect CNVs?

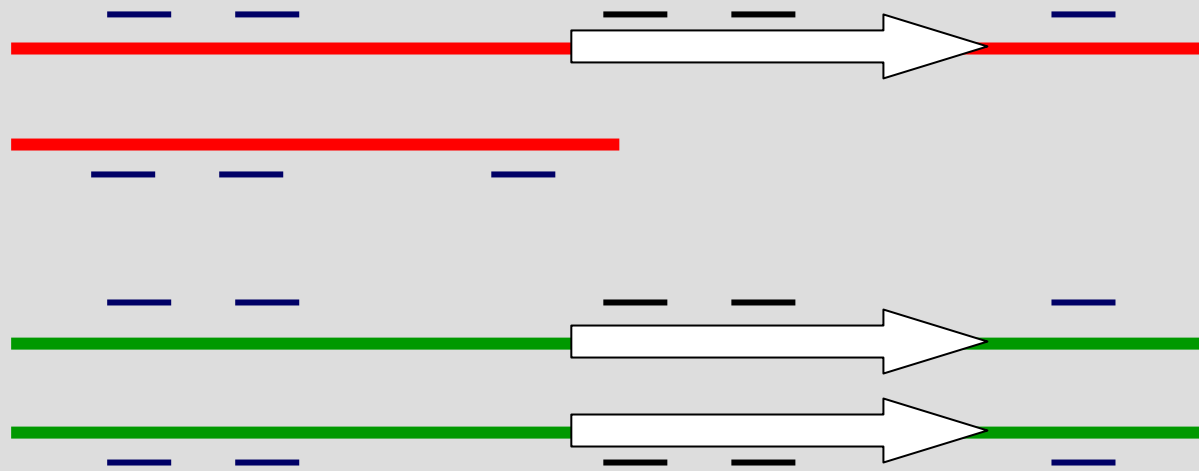
- Human-Chimp:
 - 98.9% of aligned identical
 - Only 97% aligns
 - Mostly lineage-specific deletions
- Cause of birth defects, cancer
- Blindspot for sequencing

Detect DNA by Hybridization

- *Probes* – short, single strand DNA molecules
- Apply mixture to *array* of probes, wash, photo
- Only probes that have reverse-complements light up

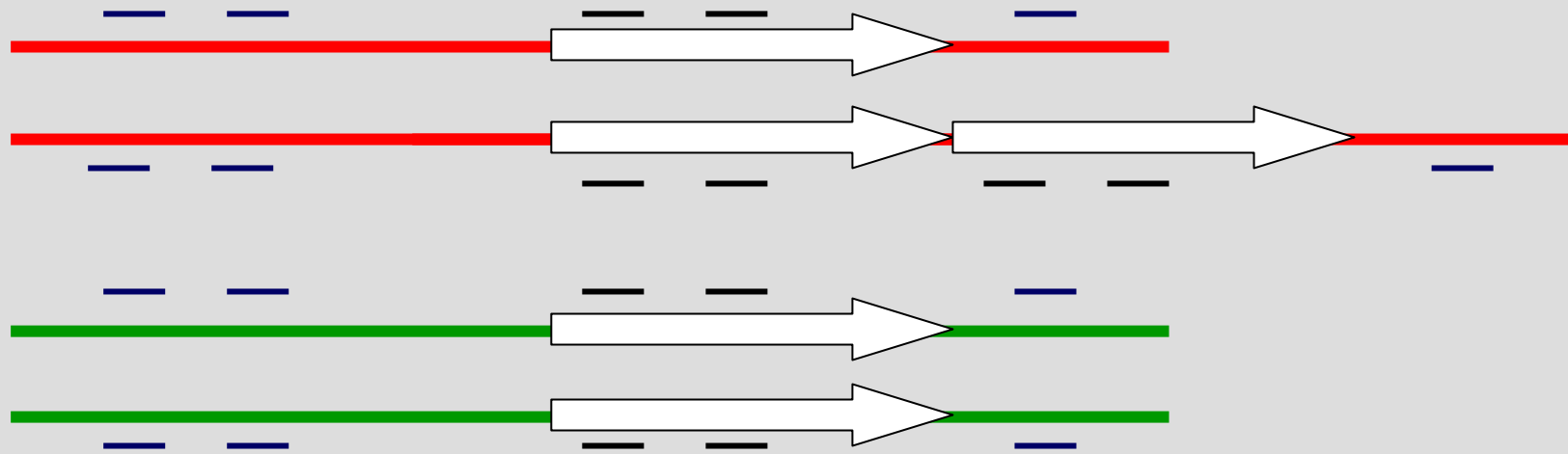


Comparative Genome Hybridization



- Hybridize different-color normal and copy-loss samples
- Probes in copy-neutral regions: **same signal**
- Probes in copy-loss regions: **2x stronger signal**

Comparative Genome Hybridization

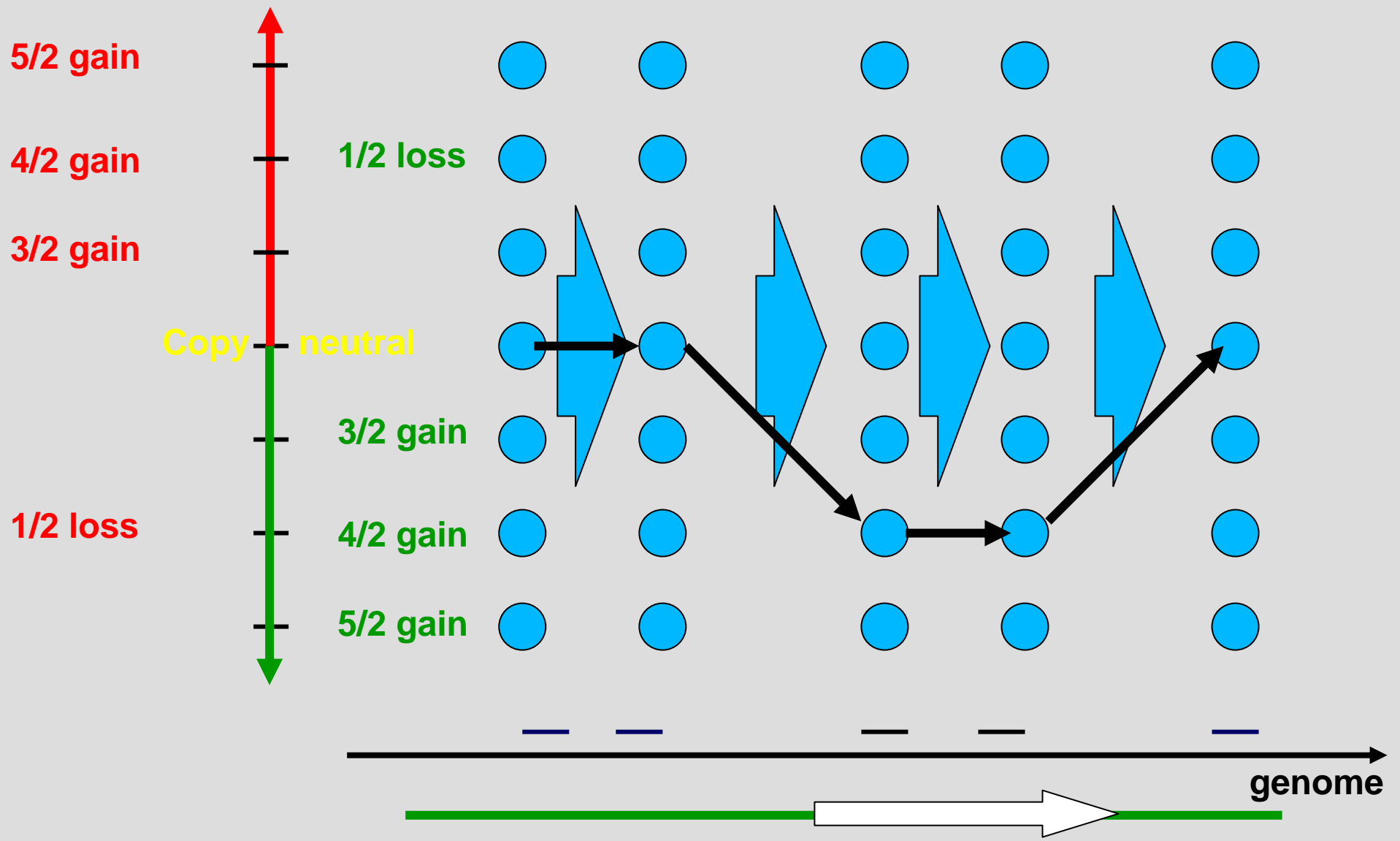


- Hybridize different-color normal and copy-loss samples
- Probes in copy-neutral regions: **same signal**
- Probes in copy-loss regions: **2x stronger signal**
- Probes in copy-gain regions: **1.5x stronger signal**

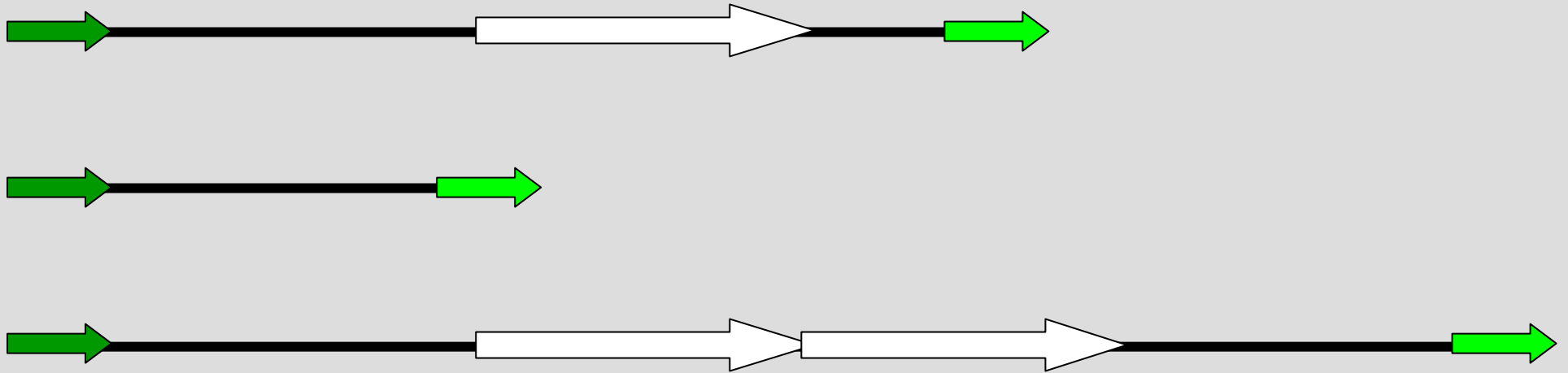
Recovering Copy#

- Input:
Probe intensities for both samples: $\{I_i\}, \{I'_i\}$
- Null:
Probes distributed $\text{Normal}(\mu_i, \sigma_i^2)$
- Alternative:
In particular segments
 $\text{Normal}(\mu_i, \sigma_i^2), \text{Normal}(\mu'_i, \sigma_i^2)$
where $\mu_i/\mu'_i \in \{k/2, 2/k\}$

Recovering Copy#



Detection of CNVs



- Alltogether: ~400 large CNVs (total > 100Mb)
- Weaknesses: medium/short, rare

Structural Variants

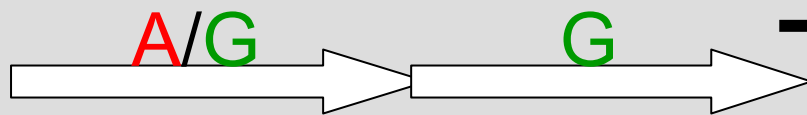
- Microsatellites
- Polymorphic inversions
- Polymorphic copy numbers
 - Direct methods
 - Indirect methods
- Somatic variants

Genotyping & Repeats

- Example:
 - Suppose typing a A/G SNP x 30 samples gives:
 - 25 samples: Het AG
 - 2 samples: No call
 - 2 samples: AA
 - 1 samples: GG
 - Not a SNP. Fixed difference between duplicons.
 - High error rate!

Genotyping & Repeats

- Fixed differences between fixed copies
- SNP in one copy:



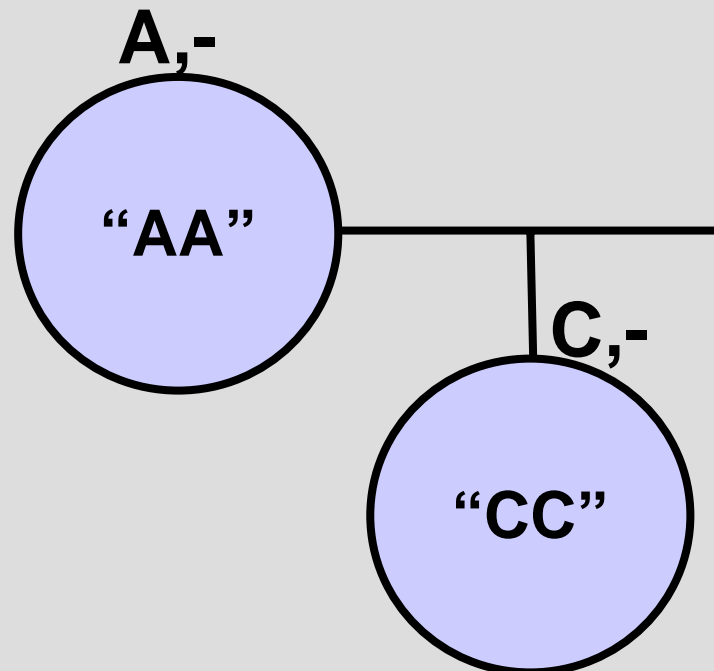
True genotype	Observed call
AA	"AA"
AG	"AG" or no-call
GG	"AG"
	"GG" - never

Genotyping & Repeats

- Fixed differences between fixed copies
- SNP in one copy
- Polymorphic copy-#
 - Can use probes for SNPs as copy# probes

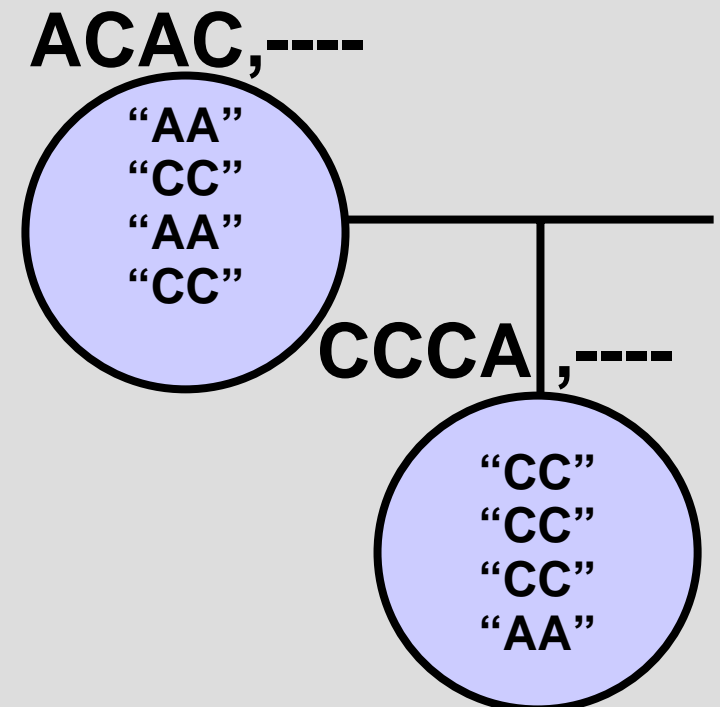
Polymorphic Deletion

- Suppose typing a A/C SNP x 30 trios gives:
 - Parents: 30 “AA”, 15 “AC”, 11 “AA”
 - 4 Parent “No Calls”, w/ homozygous kids
 - 2 offspring “No Calls”, w/ homozygous parents
 - 4 Mendel violations w/ homozygous transmissions



SNPs in Polymorphic Deletions

- Excess homozygosity
- Poor call rate
- Homozygous Mendel violations
- **Sequence of SNPs with:**
 - Similar problems
 - Same faulty transmissions and genotypes



Mining Deleted Segments

- Heuristic filtering:
 - Scan millions of SNPs for:
 - Violations with null $p < \sim 10^{-6}$
 - 2+ nearby violations with $p < \sim 10^{-2}$, same samples
- Exact scoring:
 - Likelihood of a deletion present

Likelihood of Deleted Segments

- Null: Haplotype pairs w/ errors
Use EM.

P_1	h_1 :	A	T	C	C	A
P_2	h_2 :	G	G	C	C	T
P_3	h_3 :	G	T	G	T	T
P_4	h_4 :	A	G	C	C	T
P_5	h_5 :	A	G	G	C	T

- Alternative:

P_1	h_1 :	A	T	C	C	A
P_2	h_2 :	G	G	C	C	T
P_3	h_3 :	G	T	G	T	T
P_4	h_4 :	A	G	C	C	T
P_5	h_5 :	A	G	G	C	T
P_{\cdot}	h_{\cdot} :	G	-	-	-	A

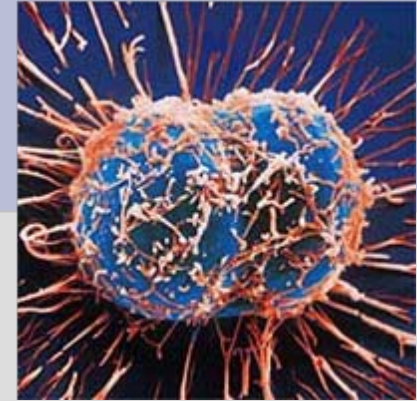
Polymorphic Human Deletions

- ~1000 deletions x $<100\text{bp}$ → ~MB
- Total:
~15Mb
- Typically
 - Tagged by a SNP/haplotype

Structural Variants

- Microsatellites
- Polymorphic inversions
- Polymorphic copy numbers
 - Direct methods
 - Indirect methods
- Somatic variants

Cancer



- Abnormal, uncontrolled division of a cell into a tumor
- Genetic illness of a cell & its descendants
- Lineage accumulates mutations and structural genetic changes
- Key functional changes:
 - + Growth
 - DNA repair
 - Mortality
 - + Blood supply
 - + Mobility

Mapping Somatic Changes

- CGH of tumor vs. normal tissue
- Detects recurrent copy# gains/losses, within/across cancer types
- Ideally: resequence.

The Cancer Genome Atlas

- Pilot: 2005-8
 - Copy numbers
 - Expression
 - Sequencing of oncogenes, tumor-suppressors...
 - Multiple technologies, cancers
- Full project:
 - Whole genome sequencing?

Summary

- Structural variants:
 - Span more than SNPs
 - More severe change
 - May have been overlooked due to technology
 - Directly relevant to cancer

Further Reading

- Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 2005 Sep 1;437(7055):69-87.
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. *Science*. 2004 Jul 23;305(5683):525-8.
- McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM; International HapMap Consortium. Common deletion polymorphisms in the human genome. *Nat Genet*. 2006 Jan;38(1):86-92.
- Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C. Copy number variation: new insights in genome diversity. *Genome Res*. 2006 Aug;16(8):949-61.
- Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet*. 2006 Jan;38(1):82-5.
- Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K. A common inversion under selection in Europeans. *Nat Genet*. 2005 Feb;37(2):129-37.
- She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*. 2004 Oct 21;431(7011):927-30.
- Eichler EE. Widening the spectrum of human genetic variation. *Nat Genet*. 2006 Jan;38(1):9-11.
- Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE. Fine-scale structural variation of the human genome.
- Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet*. 2006 Jan;38(1):75-81.

Extra Credit

1. Look at the HapMap genotypes for rs10914658 and figure out what's wrong.
2. Half a million 40kb segments are taken uniformly from 10 individuals, and their ends of 500bp are mapped onto the genome reference of a single individual. What are the chances of missing a 10%, 2kb inversion?

Project Suggestion

- Systematically look for polymorphic inversions using:
 - the HapMap data
 - Sequences from the human genome sequencing